Documentation for wyoung

wyoung is a Stata command that controls the family-wise error rate when performing multiple hypothesis tests. This document describes the algorithm employed by **wyoung** and provides simulation results that demonstrate its effectiveness in different settings. The command and its help file can be installed by typing "`ssc install wyoung, replace`" at the Stata prompt. It can also be downloaded directly from:
https://github.com/reifjulian/wyoung

After installation, type "`help wyoung`" at the Stata prompt to view examples and learn the syntax. Companion Stata code for the numerical simulations we describe below can be downloaded from:
www.nber.org/workplacewellness/s/wyoung_simulations.do.

The Stata command **wyoung** was developed for use in the Illinois Workplace Wellness Study. Please cite the command as:

Jones, D., D. Molitor, and J. Reif. "What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study." *Quarterly Journal of Economics*, November 2019, 134(4): 1747-1791.

1 Methodology

Multiple hypotheses arise when there are multiple outcomes of interest, multiple subgroups of interest, multiple independent variables of interest, or some combination thereof. Consider testing $K > 1$ different null hypotheses. The family-wise error rate (FWER) is the probability of rejecting at least one true null hypothesis (i.e., a "false discovery") belonging to this "family" of K hypotheses. A procedure is said to provide strong control of the FWER if it does not depend on which of the K null hypotheses happen to be true.

We estimate the FWER using the free step-down resampling method of Westfall and Young (1993) (Algorithm 2.8, p. 66–67). The procedure consists of the following steps:

1. Estimate $\{\widehat{\beta}_1, \widehat{\beta}_2, ..., \widehat{\beta}_K\}$. Estimate the conventional, unadjusted p-values $\{p_1, p_2, ..., p_K\}$ that correspond to separately testing each null hypothesis $\widehat{\beta}_k = 0$. Without loss of generality, assume the estimated p-values are indexed such that $p_1 \leq p_2 \leq ... \leq p_K$.

2. Draw with replacement from the dataset to create a bootstrap sample.

   (a) Estimate $\{\widehat{\beta}^*_{i1}, \widehat{\beta}^*_{i2}, ..., \widehat{\beta}^*_{iK}\}$. Estimate the conventional, unadjusted p-values $\{p^*_{i1}, p^*_{i2}, ..., p^*_{iK}\}$ that correspond to separately testing each null hypothesis $\widehat{\beta}^*_{ik} = \widehat{\beta}_k$. The k index here corresponds to the ranking computed in step 1. It will not generally be the case that $p^*_{i1} \leq p^*_{i2} \leq ... \leq p^*_{iK}$.

   (b) Enforce monotonicity with respect to the original ordering in step 1 by computing the successive minima:

$$\begin{aligned} q^*_{iK} &= p^*_{iK} \\ q^*_{i,K-1} &= \min(q^*_{iK}, p^*_{i,K-1}) \\ q^*_{i,K-2} &= \min(q^*_{i,K-1}, p^*_{i,K-2}) \\ &\vdots \\ q^*_{i1} &= \min(q^*_{i2}, p^*_{i1}). \end{aligned}$$

3. Repeat step 2 N times. For each bootstrap sample i and hypothesis k, define the indicator $COUNT_{ik} = 1$ if $q^*_{ik} \leq p_k$ and 0 otherwise.[1]

4. For each hypothesis $k = 1, 2, ..., K$, calculate the fraction of successive minima that were lower than the original p-value:

$$r_k = \frac{1}{N} \sum_{i=1}^{N} COUNT_{ik}.$$

5. Enforce monotonicity using successive maximization to calculate the adjusted p-value:

$$\begin{aligned} p^{adj}_1 &= r_1 \\ p^{adj}_2 &= \max(r_1, r_2) \\ &\vdots \\ p^{adj}_K &= \max(r_{K-1}, r_K). \end{aligned}$$

[1] To compute "single-step" p-values instead of "step-down" p-values, define the indicator $COUNT_{ik} = 1$ if $\min\{p^*_{i1}, p^*_{i2}, ..., p^*_{iK}\} < p_k$ and 0 otherwise. Resampling-based single-step methods often control family-wise type III (sign) error rates. Whether their step-down counterparts also control type III error rates is unknown (Westfall and Young, 1993, p. 51).

This resampling algorithm exhibits strong control of the FWER under subset pivotality, which is a multivariate generalization of pivotality.[2] This condition requires that the multivariate distribution of any subvector of p-values is unaffected by the truth or falsehood of hypotheses corresponding to p-values not included in the subvector. The condition is satisfied in many settings, including testing the significance of coefficients in a general multivariate regression model with possibly non-normal or heteroskedastic errors (Westfall and Young, 1993, p. 122–123).

It is possible for this algorithm to produce adjusted p-values that are smaller than unadjusted p-values. For example, consider the extreme case where the number of bootstraps is equal to one (so that $N = 1$ in steps 3 and 4). Then all adjusted p-values are equal to either zero or one. The ones that are equal to zero will, of course, be smaller than the unadjusted values. For this reason, we recommend employing a large number of bootstraps. Westfall and Young (1993) recommend at least 10,000 bootstrap draws. If adjusted p-values remain significantly smaller than the unadjusted p-values, even when the number of bootstraps is large, this may indicate model misspecification. For example, in simulations with clustered errors (described below), we found that adjusted p-values are frequently smaller than unadjusted values when we fail to employ a cluster bootstrap.

2 Simulations

We ran simulations to evaluate the effectiveness and statistical power of the resampling algorithm described above. Let μ be a ten-dimensional zero vector $(0, 0, ..., 0)'$. Let I be a 10×10 identity matrix. Let Σ be a 10×10 covariance matrix where all off-diagonal elements are equal to 0.9. The data-generating process for each simulation scenario is described below:

1. Normal i.i.d. errors (ten outcomes)

   $e \sim \mathcal{N}(\mu, I)$

   $Y = e$

2. Normal i.i.d. errors (one outcome, ten subgroups)

   $e \sim \mathcal{N}(0, 1)$

   $Y = e$

3. Correlated errors (ten outcomes)

   $X \sim \mathcal{N}(\mu, I)$

   $e \sim \mathcal{N}(\mu, \Sigma)$

   $Y = 0.2X + e$

4. Lognormal, mean-zero i.i.d. errors (ten outcomes)[3]

[2] The sampling distribution of a pivotal statistic does not depend upon which distribution generated the data; the t-statistic is a common example.

[3] The mean of the standard lognormal distribution is $\sqrt{\exp[1]}$.

$$e \sim \exp[\mathcal{N}(\mu, I)] - \sqrt{\exp[1]}$$
$$Y = e$$

We simulated 2,000 datasets for each of these four data-generating processes. In each of these 2,000 simulations, we estimated a series of 10 regressions:

$$Y_i = \alpha + \beta_i X_i + \varepsilon_i, i = 1...10.$$

The sample size for each regression was 100. The regressor $X_i \sim N(0, 1)$ in simulations 1, 2, and 3. In scenario 4, the regressor is just a constant equal to one (α is omitted). There are ten null hypotheses that correspond to these ten regressions: $\beta_i = 0, i = 1, ..., 10$. These ten null hypotheses are all true in scenarios 1, 2, and 4; the hypotheses are all false in scenario 3 (correlated errors).

Table 1 compares the effectiveness of the Westfall-Young resampling algorithm to other well-known multiple inference adjustment methods.[4] Each column in the table reports how often at least one null hypothesis was rejected using each adjustment method. When outcomes are independent and normally distributed, the probability that at least one of the ten hypotheses is statistically significant is equal to $1 - (1 - 0.05)^{10} = 0.401$. This calculation accords well with the simulation: the first row of column (1) reports that at least one of the ten hypotheses was rejected at $\alpha = 0.05$ in 39.8 percent of the 2,000 simulations when no adjustment was performed. By contrast, the Bonferroni-Holm, Sidak-Holm, and Westfall-Young adjustments reject at least one null hypothesis only about 4 percent of the time, thus achieving a FWER of less than 5 percent.

In column (2), the ten hypotheses arise from examining multiple subgroups rather than multiple outcome variables. Failing to adjust the p-values again results in a high rejection rate of nearly 40 percent. The Bonferroni-Holm, Sidak-Holm, and Westfall-Young adjustment methods, however, all achieve rejection rates of around 5 percent.

The downside of the Bonferroni-Holm and Sidak-Holm adjustment methods is that they assume outcomes are independent and therefore can be too conservative when outcomes are correlated. This is demonstrated in column (3), which reports rejection rates for a scenario where the ten null hypotheses are all false. Here, the Bonferroni-Holm and Sidak-Holm methods reject at least one hypothesis only about 35 percent of the time. The Westfall-Young resampling algorithm, however, achieves a rejection rate in excess of 50 percent.

Although traditional adjustment methods such as Bonferroni-Holm and Sidak-Holm are generally thought to be conservative, Westfall and Young (1993) emphasize that these traditional methods can actually over-reject when the data-generating process is nonnormal. This is demonstrated in column (4): the resampling method of Westfall-Young achieves a FWER of under 6 percent, but the Bonferroni-Holm and Sidak-Holm methods reject at least one null hypothesis over 20 percent of the time.

[4]The Bonferroni-Holm and Sidak-Holm (step-down) p-values are calculated as follows. Sort the K unadjusted p-values so that $p_1 \leq p_2 \leq ... \leq p_K$. The Bonferroni-Holm adjusted p-values are calculated as $\{p_1 K, \max[p_1, p_2(K-1)], ..., \max[p_{K-1}, p_K]\}$. The Sidak-Holm adjusted p-values are calculated as $\{1-(1-p_1)^K, \max[p_1, 1-(1-p_2)^{(K-1)}], ..., \max[p_{K-1}, p_K]\}$. If the calculation yields a value larger than one, then the adjusted p-value is set equal to one.

2.1 Clustered Standard Errors

Westfall and Young (1993) do not discuss how to perform multiple inference in regression models where observations can be grouped into clusters, with model errors correlated within clusters. The presence of clustered errors does not violate subset pivotality, which is automatically satisfied in linear regression models. However, in this case it is important that the resampling in step 2 of the procedure be done over entire clusters rather than individual observations. This is accomplished by specifying the **cluster()** option of the **wyoung** command.

We demonstrate the importance of resampling over clusters by performing another set of simulations. Again, let μ be a ten-dimensional zero vector $(0, 0, ..., 0)'$, and let I be a 10×10 identity matrix. The data-generating process for this simulation scenario is

5. Serially correlated errors (ten outcomes)

   $i = 1...100$ clusters

   $t = 1...10$ time periods

   $\eta_i \sim \mathcal{N}(\mu, I)$

   $e_{it} \sim \mathcal{N}(\mu, I)$

   $Y_{it} = \eta_i + e_{it}$

We again simulated 2,000 datasets. In each simulation, we estimated the following ten regressions:

$$Y_{it} = \alpha + \beta_i D_{it} + \varepsilon_{it}, i = 1...10,$$

where the dummy variable $D_{it} = 1\{t > START_i\}$ and $START_i$ is a Poisson random variable with mean equal to five. We estimated these regressions under two different assumptions about the standard errors (homoskedastic or clustered) and with and without a bootstrap cluster. Our results are reported in Table 2.

Comparing column (2) to column (1) in the first row of Table 2 shows that estimating the model using clustered standard errors results in a smaller FWER relative to a model that assumes errors are homoskedastic. Nevertheless, the rejection rate for the unadjusted value in column (2) still significantly exceeds 5 percent because this specification does not account for the number of hypotheses being tested.[5]

The second and third rows of Table 2 show that the Bonferroni-Holm and Sidak-Holm corrections achieve a 5 percent rejection rate when the standard errors are clustered. This is unsurprising since the outcome variables in this simulation are independent.

The fourth row of Table 2 demonstrates the importance of properly accounting for clustered standard errors when implementing the Westfall-Young correction. Column (2) shows that (erroneously) employing a simple bootstrap that resamples over individual observations causes the Westfall-Young correction to perform worse than even the unadjusted specification. However, column (3) shows that the Westfall-Young correction achieves a 5 percent rejection rate when the cluster bootstrap is employed.

[5]By construction, the values in columns (2) and (3) are identical in the first three rows because these two columns vary only the bootstrapping methodology, which matters only for the Westfall-Young correction.

2.2 Linear and nonlinear combinations

wyoung allows researchers to perform multiple inference when testing a hypothesis about any linear or nonlinear combination of coefficients. We used simulated data to investigate the effectiveness of using **wyoung** to test linear and nonlinear restrictions of two linear regression coefficients. Let μ be a ten-dimensional zero vector $(0, 0, ..., 0)'$. Let I be a 10×10 identity matrix. The data-generating process is:

6. Multiple restrictions (ten outcomes)

   $X_1 \sim \mathcal{N}(\mu, I)$

   $X_2 \sim \mathcal{N}(\mu, I)$

   $e \sim \mathcal{N}(\mu, I)$

   $Y = 2X_1 + 0.5X_2 + e$

We simulated 2,000 datasets using this data-generating process. In each simulation, we estimated a series of 10 regressions:

$$Y_i = \alpha + \beta_{i1}X_{i1} + \beta_{i2}X_{i2} + \varepsilon_i, i = 1...10.$$

The sample size for each regression was 100. We tested the following null hypotheses: (1) the linear restriction $\beta_1 - 4\beta_2 = 0$; and (2) the nonlinear restriction $\beta_1\beta_2 - 1 = 0$. Both these null hypotheses are true by construction.

Rejection rates, reported in Table 3, exceed forty percent if no adjustment is performed. By contrast, the rejection rates corresponding to the adjusted p-values are about five percent when testing the linear restriction and about six percent when testing the nonlinear restriction.

References

Westfall, Peter H and S Stanley Young, *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279 (John Wiley & Sons, 1993).

Table 1: Family-wise rejection proportions at $\alpha = 0.05$

	(1)	(2)	(3)	(4)
Adjustment method	Normal errors	Multiple subgroups	Correlated errors	Lognormal errors
Unadjusted	0.398	0.387	0.685	0.577
Bonferroni-Holm	0.040	0.047	0.344	0.234
Sidak-Holm	0.040	0.051	0.347	0.237
Westfall-Young	0.041	0.045	0.513	0.058
Num. observations	100	100	100	100
Num. hypotheses	10	10	10	10
Hypotheses are true	Y	Y	N	Y

Notes: Table reports the fraction of 2,000 simulations where at least one null hypothesis in a family of 10 hypotheses was rejected. All hypotheses are true for the simulations reported in columns (1), (2), and (4), i.e., lower rejection rates are better. All hypotheses are false for the simulation reported in column (3), i.e., higher rejection rates are better. The Westfall-Young correction is performed using 1,000 bootstraps.

Table 2: Family-wise rejection proportions at $\alpha = 0.05$, when the data generating process is serially correlated

	(1)	(2)	(3)
Unadjusted	0.652	0.401	0.401
Bonferroni-Holm	0.187	0.049	0.049
Sidak-Holm	0.188	0.049	0.049
Westfall-Young	0.191	0.498	0.046
Num. observations	1,000	1,000	1,000
Num. hypotheses	10	10	10
Model std. errors	Homoskedastic	Clustered	Clustered
Cluster bootstrap	N	N	Y

Notes: Table reports the fraction of 2,000 simulations where at least one null hypothesis in a family of 10 hypotheses was rejected. The difference between columns (1) and (2) is the assumption about the standard errors (homoskedastic or clustered). The difference between columns (2) and (3) is the method of bootstrapping (resampling over individual observations versus clusters), which matters only for the Westfall-Young correction. All null hypotheses are true, i.e., lower rejection rates are better. Each simulation generated 100 panels (clusters) with 10 time periods. The Westfall-Young correction is performed using 1,000 bootstraps.

Table 3: Family-wise rejection proportions at $\alpha = 0.05$, when testing hypotheses with multiple restrictions

	(1)	(2)
Adjustment method	Linear restriction	Nonlinear restriction
Unadjusted	0.440	0.435
Bonferroni-Holm	0.052	0.064
Sidak-Holm	0.052	0.066
Westfall-Young	0.051	0.063
Num. observations	100	100
Num. hypotheses	10	10

Notes: Table reports the fraction of 2,000 simulations where at least one null hypothesis in a family of 10 hypotheses was rejected. All null hypotheses are true, i.e., lower rejection rates are better. The Westfall-Young correction is performed using 1,000 bootstraps.